Имя – это feature

Виталий Худобахшов, 2017

SNA как он есть

Анализа 68 лайков в Facebook достаточно, чтобы определить цвет кожи испытуемого (с 95% вероятностью), его гомосексуальность (88% вероятности) и приверженность Демократической или Республиканской партии (с 85% вероятностью)

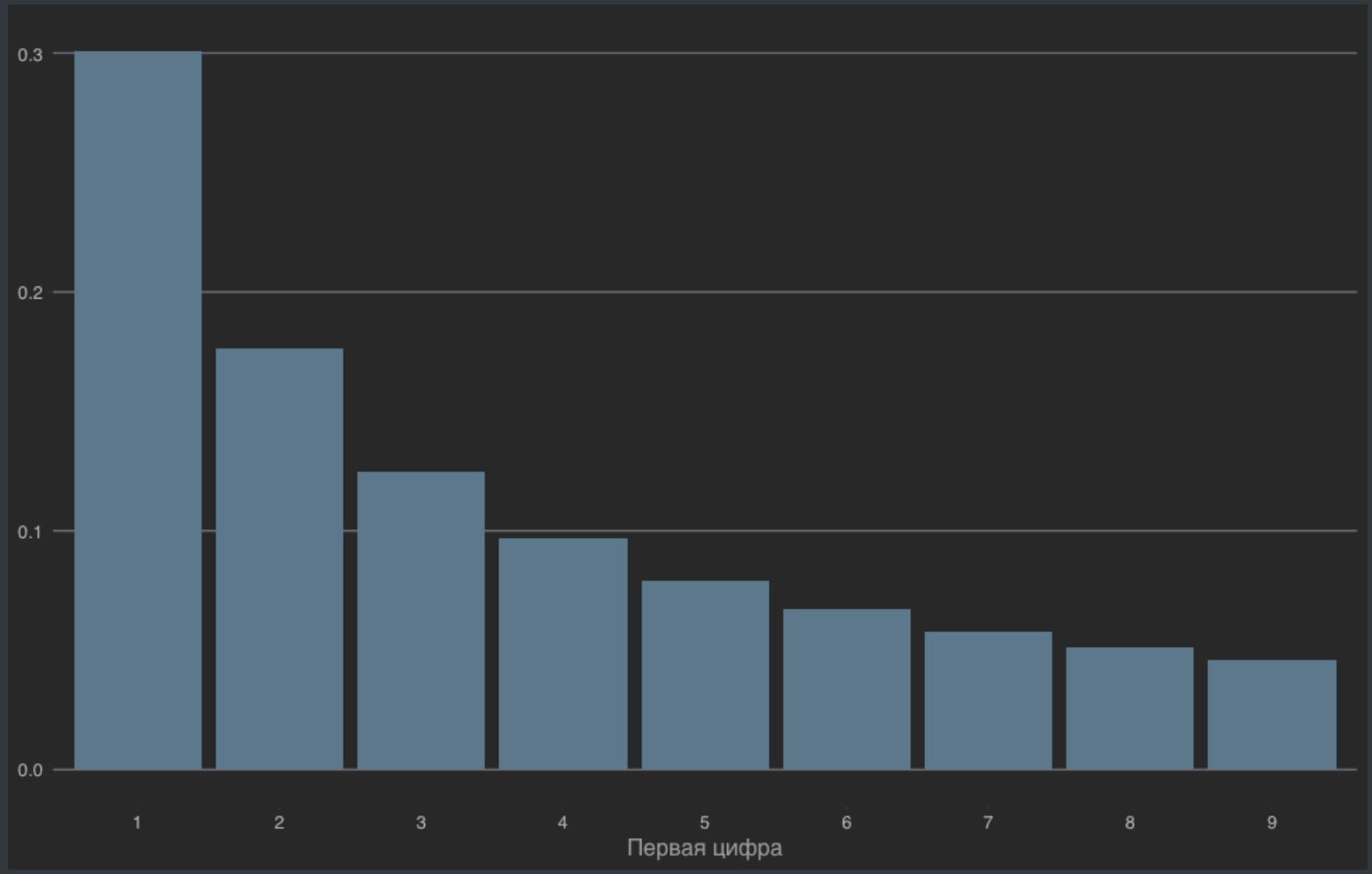
Михал Козински

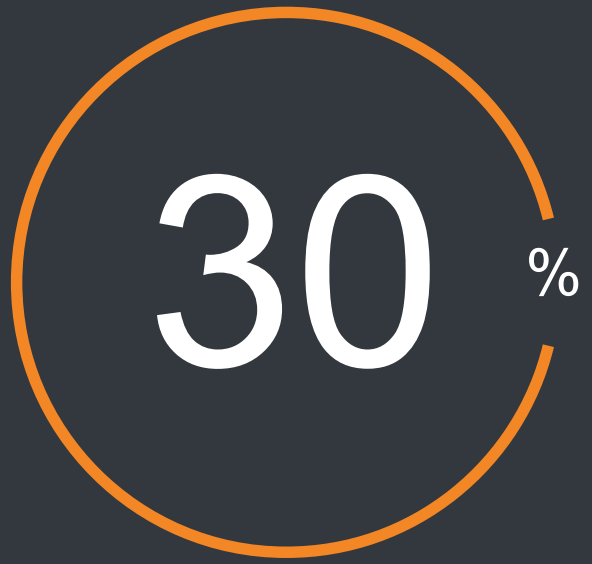
Источник: M. Kosinski et al. Private traits and attributes are predictable from digital records of human behavior, 2013

Необычные распределения

- Зайдите в какой-нибудь каталог, где много файлов
- Какая доля файлов имеет размер, начинающийся с 9?
  - Ожидание: 11%
  - Реальность: 5%
- Какая доля файлов имеет размер, начинающийся с 1?
  - Ожидание: 11%
  - Реальность: 30%

Закон Бенфорда





Первая цифра = 1

Верите ли вы что?

Вероятность быть сбитым автомобилем, если вас зовут Алексей, выше, чем если вас зовут Василий.

А если так?

У россиян с фамилиями Орлов, Кузьмин и Виноградов самый высокий уровень просрочки выплат по микрокредитам.

Life.ru 16 октября 2017 (goo.gl/tV8WzA)

А в это вы верите?

Девушки с именем Кира более одиноки, чем с именем Наташа.

История с Tinder

Исследовательский процесс

• Этого не может быть	отрицание
• Нет, я опять это вижу	гнев
• Tinder знает что-то, чего не знаю я	торг
• Это все довольно грустно	депрессия
• Но что тут поделаешь	принятие

Гипотеза №1

Вероятность быть одиноким зависит от имени.

Статусы в социальных сетях

ВКонтакте

- Не замужем (single)
- Замужем (married)
- Влюблена (love)
- В отношениях (relationship)

Одноклассники

- Замужем (married)
- Влюблена (love)

Как определить «одиночество»?

ВКонтакте

- $одиночество_1 = (\#single + \#searching) / \#all$
- $одиночество_2 = 1 – (\#married + \#love + \#relationship) / \#all$
- $одиночество_{вк} = (одиночество_1 + одиночество_2) / 2$

Одноклассники

- $одиночество_{ок} = 1 - (\#married + \#love) / \#all$

Список имен

Высокочастотные

- Анастасия
- Екатерина
- Елена
- Мария
- Наталья

Среднечастотные

- Дарья
- Алина
- Ксения
- Александра

Низкочастотные

- Кира
- Инесса
- Лейла

Девушки 20-35

Имя	Одиночество
Кира	0.524
Алина	0.479
Лейла	0.477
Александра	0.458
Дарья	0.446
Мария	0.444
Анастасия	0.439
Екатерина	0.421
Ксения	0.419
Инесса	0.404
Елена	0.398
Наталья	0.389
Татьяна	0.384

Группы имен

Верхняя

Средняя

Нижняя

Имя	Одиночество
Лейла	0.962
Кира	0.948
Алина	0.942
Дарья	0.942
Александра	0.933
Мария	0.932
Ксения	0.927
Анастасия	0.924
Екатерина	0.921
Инесса	0.920
Елена	0.917
Наталья	0.917
Татьяна	0.911

Возможные объяснения

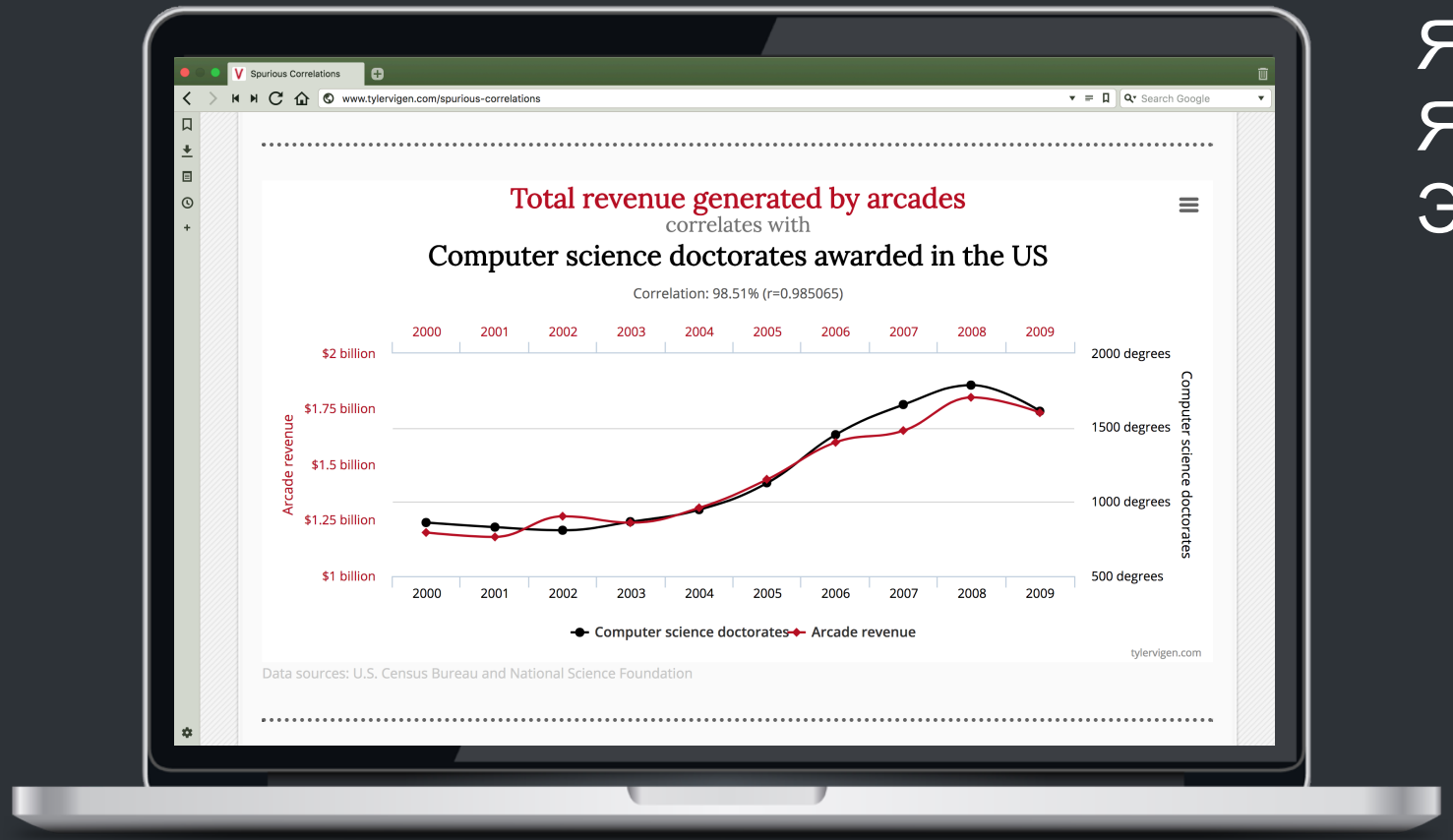
Конструктивные

- Влияние частотности и популярности имен
- География
- Влияние ботов и спамеров
- Социальные факторы

Неконструктивные

- Статистически незначимые результаты
- Ложная корреляция

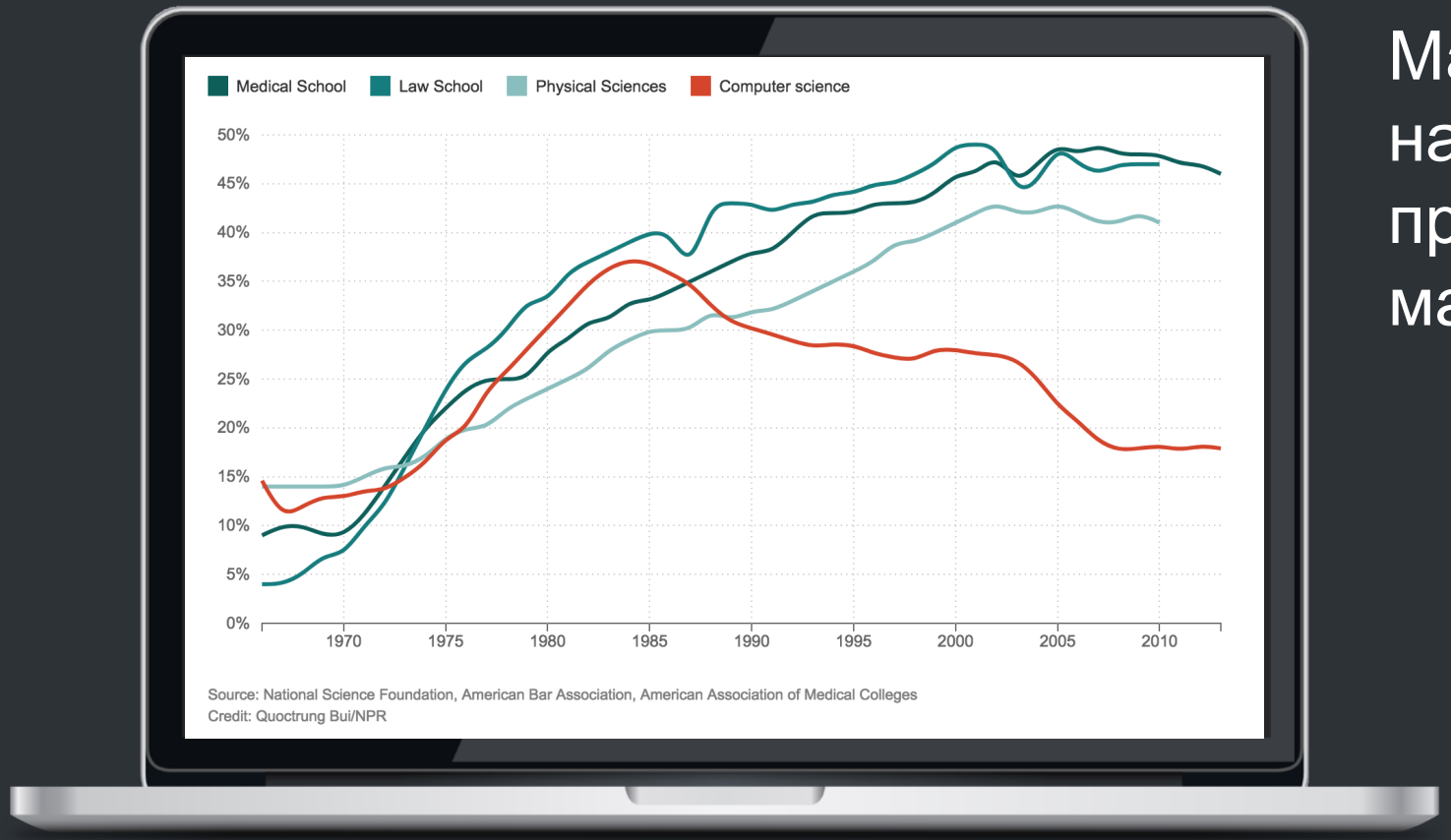
Самое простое решение



Я – data scientist.
Я не хочу ничего решать,
Это все ложная корреляция!

Источник: www.tylervigen.com/spurious-correlations

Доля женщин в соответствующей специальности



Маркетинг компьютерных игр направлен преимущественно на мальчиков

Источник: npr.org (goo.gl/tpjNUs)

Статистическая значимость

Качественные критерии

- Разная аудитория
- Различный набор статусов
- Различный способ установки статуса
- Две системы случайно дают одинаковый результат?

Статистические критерии

- Тест Манна-Уитни для перемешанных 100 имен дает p-value = 0.00026

Популярность имен

20-35 лет	28 лет	22 года
Кира	Кира	Кира
Алина	Лейла	Лейла
Лейла	Алина	Алина
Александра	Александра	Александра
Дарья	Мария	Мария
Мария	Анастасия	Екатерина
Анастасия	Дарья	Дарья
Екатерина	Ксения	Ксения
Ксения	Елена	Анастасия
Инесса	Екатерина	Елена
Елена	Инесса	Инесса
Наталья	Наталья	Наталья
Татьяна	Татьяна	Татьяна

География

Москва	Екатеринбург	Санкт-Петербург	Новосибирск
Кира	Кира	Кира	Кира
Алина	Алина	Алина	Лейла
Лейла	Лейла	Лейла	Алина
Александра	Александра	Александра	Александра
Инесса	Мария	Дарья	Дарья
Дарья	Дарья	Анастасия	Мария
Мария	Инесса	Мария	Анастасия
Анастасия	Анастасия	Ксения	Ксения
Ксения	Ксения	Инесса	Екатерина
Екатерина	Екатерина	Екатерина	Инесса
Елена	Елена	Елена	Елена
Наталья	Наталья	Наталья	Наталья
Татьяна	Татьяна	Татьяна	Татьяна

Нужна вспомогательная гипотеза

Скорее всего различия есть не только в степени одиночества, но и еще где-то

Образование

Гипотеза №2

А что если более образованные люди более одиноки?

Как оценить IQ?

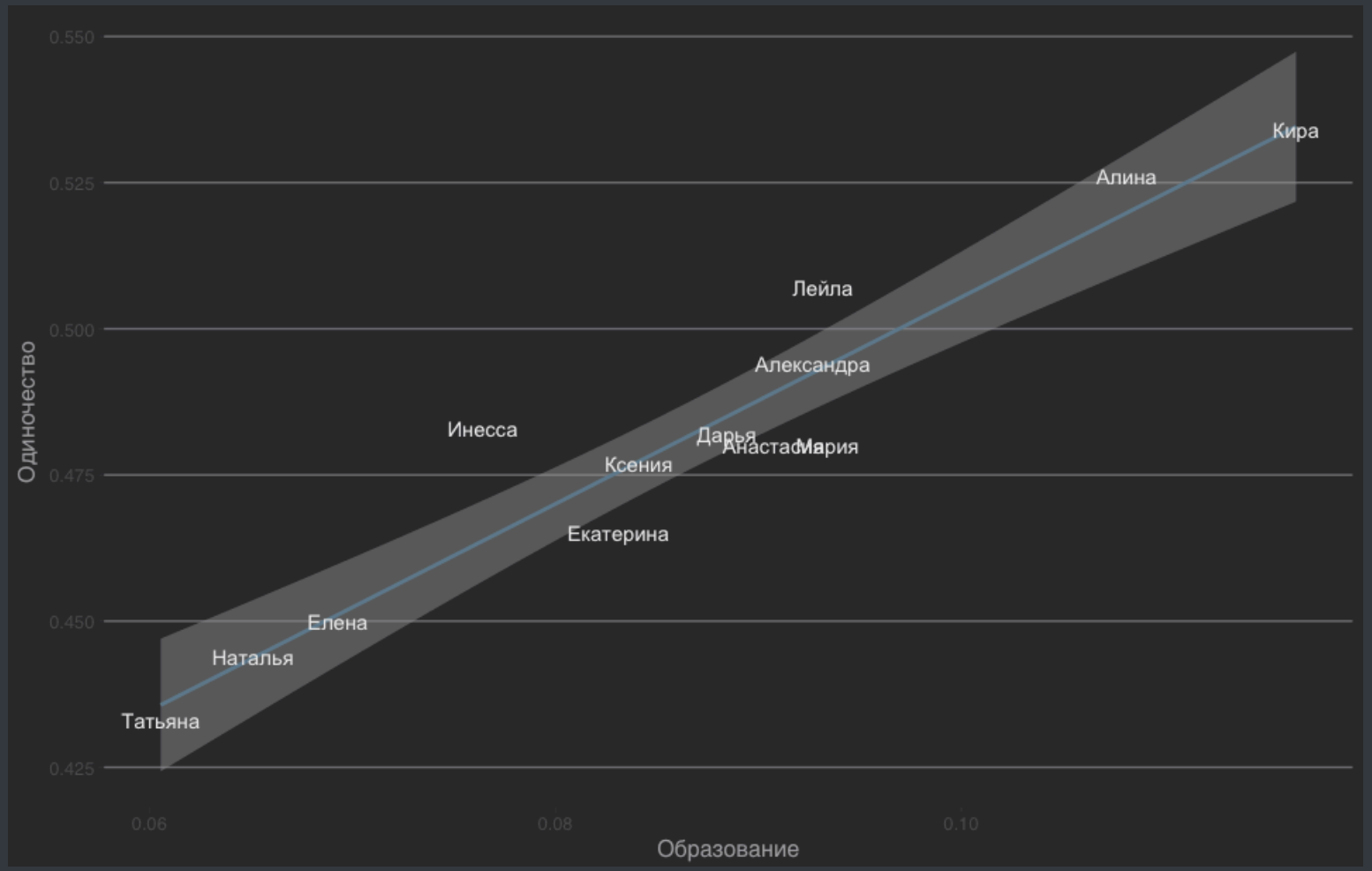
- Можно быть как М. Козински (тесты, приложения и пр.)
- Нам не нужен IQ конкретного человека, нам нужна качественная оценка для группы людей
- Хорошее образование – хорошее приближение

Как оценить образование группы людей?

- Выберем несколько ведущих университетов
- Пусть *для каждого* имени p = #school / #all, доля людей в данном университете ко всем в данном городе
- Тогда p – оценка IQ для людей с таким именем

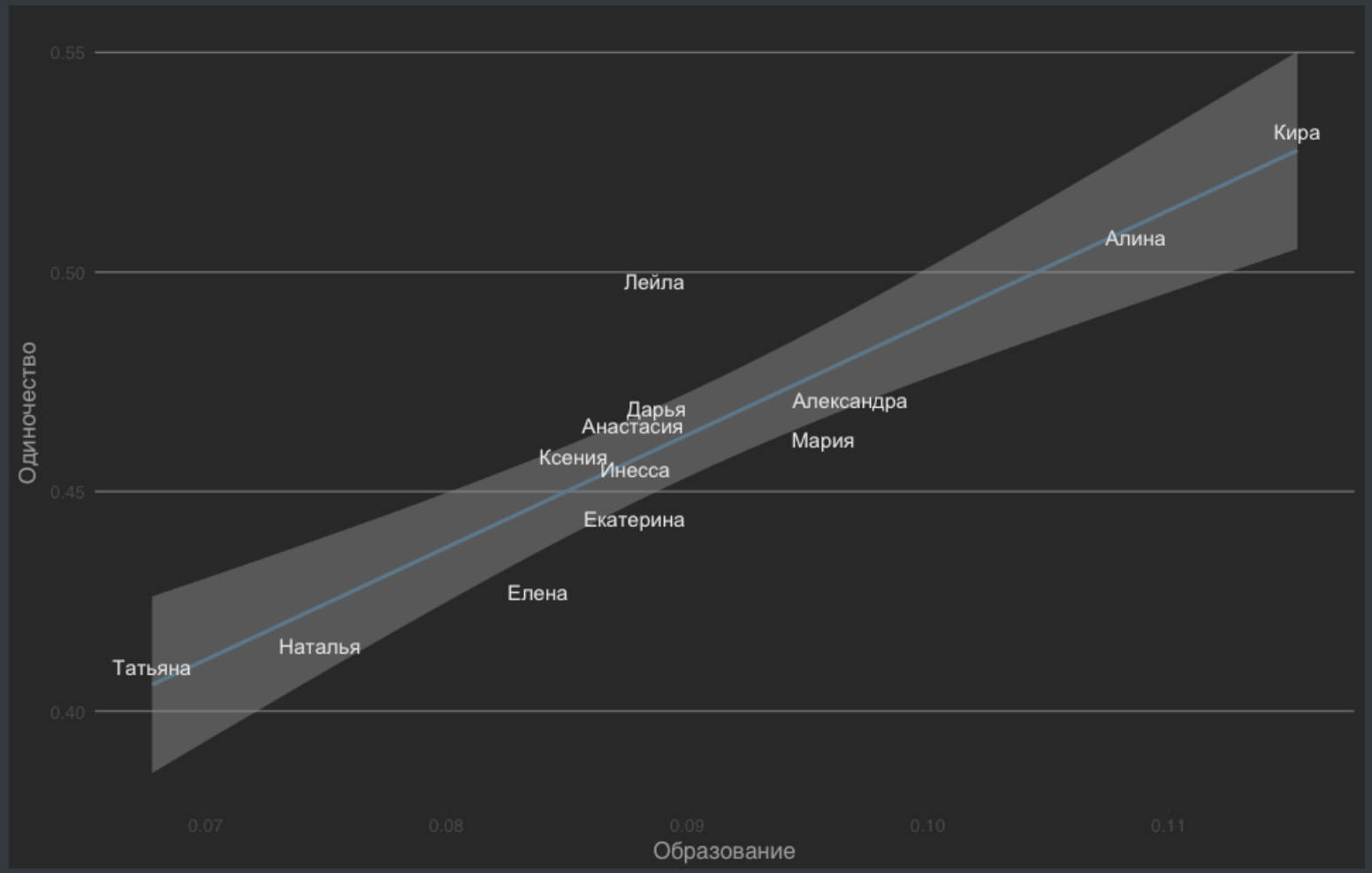
Статистика для МГУ

Одиночество	Образование
Кира	Кира
Алина	Алина
Лейла	Мария
Александра	Лейла
Дарья	Александра
Мария	Анастасия
Анастасия	Дарья
Екатерина	Ксения
Ксения	Екатерина
Инесса	Инесса
Елена	Елена
Наталья	Наталья
Татьяна	Татьяна



Статистика для СПбГУ

Одиночество	Образование
Кира	Кира
Алина	Алина
Лейла	Александра
Александра	Мария
Дарья	Дарья
Мария	Лейла
Анастасия	Инесса
Екатерина	Екатерина
Ксения	Анастасия
Инесса	Ксения
Елена	Елена
Наталья	Наталья
Татьяна	Татьяна



А что если посчитать значение «одиночества» для университета целиком?

Рейтинг университетов

Наш рейтинг

Университет	Одиночество
МГУ	0.536
СПбГУ	0.510
ВШЭ (Москва)	0.498
НГУ	0.452
МФТИ	0.446
ИТМО	0.432
Политех	0.406

Национальный рейтинг

Университет	Рейтинг 2017
МГУ	1
СПбГУ	3
ВШЭ (Москва)	4
НГУ	5
МФТИ	6
ИТМО	7
Политех	14

Источник: www.univer-rating.ru

Бот или не бот?

Понятно, что ботов довольно много, но как именно они влияют на распределение?

Распределение имен заблокированных аккаунтов

Основная выборка

- Кира / Екатерина = 0.02
- Дарья / Екатерина = 0.35

Заблокированная выборка

- Кира / Екатерина = 0.03
- Дарья / Екатерина = 0.30

Распределение в основной выборке и в выборке заблокированных пользователей совпадает.

Если бы к такому распределению приводили боты, то среди них Киры должны встречаться в 3 раза чаще.

Так в чем же причина?

- Если взять пары (мать, дочь) в соц. графе
- Количество пар убывает с увеличением |R(мать)-R(дочь)|, где R – это ранг в таблице
- Это значит, что имена циркулируют внутри социальной группы

Закон Бенфорда для имен

- Если взять около 100 случайных женских имен
- И посмотреть как распределены первые цифры количества людей с таким именем

Первая цифра = 1

А как из этого извлечь хоть какую-то пользу?

Насколько люди легко расстаются с деньгами

Имя	Одиночество
Кира	0.524
Алина	0.479
Лейла	0.477
Александра	0.458
Дарья	0.446
Мария	0.444
Анастасия	0.439
Екатерина	0.421
Ксения	0.419
Инесса	0.404
Елена	0.398
Наталья	0.389
Татьяна	0.384

Имя	Деньги
Кира	8.0
Алина	16.5
Лейла	19.0
Александра	19.4
Дарья	19.4
Анастасия	19.5
Ксения	19.7
Мария	20.1
Екатерина	21.8
Елена	24.4
Татьяна	28.9
Наталья	29.2
Инесса	31.8

Насколько люди легко расстаются с деньгами

Имя	Одиночество
Марк	0.541
Никита	0.523
Филипп	0.520
Кирилл	0.515
Василий	0.511
Дамир	0.508
Григорий	0.508
Иван	0.504
Андрей	0.502
Александр	0.498
Алексей	0.496
Сергей	0.489
Виталий	0.480

Имя	Деньги
Марк	8.0
Филипп	16.5
Кирилл	19.0
Никита	19.4
Дамир	19.4
Андрей	19.5
Алексей	19.7
Иван	20.1
Василий	21.8
Александр	24.4
Сергей	28.9
Григорий	29.2
Виталий	31.8

Выводы

- В социальных сетях можно говорить об устойчивых группах имен и естественном упорядочивании имен согласно ряду свойств
- Замена одного свойства на другое сохраняет группы
- Различия в свойствах имен обусловлено социальными факторами

Полезные ссылки

- Про связь имени и лица http://www.npr.org/sections/health-shots/2017/02/27/517496915/your-name-might-shape-your-face-researchers-say
- Data Science: Про любовь, имена и не только (Часть I) https://habrahabr.ru/company/odnoklassniki/blog/336390/
- Data Science: Про любовь, имена и не только (Часть II) https://habrahabr.ru/company/odnoklassniki/blog/337368/
- Лаборатория анализа данных http://insideok.ru/dsl


одноклассники